JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 6, Number 1, 1999
Mary Ann Liebert, Inc.
Pp. 37–51

Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data

RICHARD DESPER,[1] FENG JIANG,[2] OLLI-P. KALLIONIEMI,[3] HOLGER MOCH,[2]
CHRISTOS H. PAPADIMITRIOU,[4] and ALEJANDRO A. SCHÄFFER[5]

ABSTRACT

Comparative genome hybridization (CGH) is a laboratory method to measure gains and losses of chromosomal regions in tumor cells. It is believed that DNA gains and losses in tumor cells do not occur entirely at random, but partly through some flow of causality. Models that relate tumor progression to the occurrence of DNA gains and losses could be very useful in hunting cancer genes and in cancer diagnosis. We lay some mathematical foundations for inferring a model of tumor progression from a CGH data set. We consider a class of tree models that are more general than a path model that has been developed for colorectal cancer. We derive a tree model inference algorithm based on the idea of a maximum-weight branching in a graph, and we show that under plausible assumptions our algorithm infers the correct tree. We have implemented our methods in software, and we illustrate with a CGH data set for renal cancer.

Key words: algorithms, branching, cancer genetics, comparative genome hybridization, renal cancer, tree inference

1. INTRODUCTION

CANCER IS ASSOCIATED with a sequence of genetic changes that cause the cell cycle division, cell differentiation, or cell death processes to go out of control. Tremendous advances in molecular biology have enabled researchers to measure genetic changes in tumor cells, but it is still very difficult to distinguish causes from effects. Cancer genes can be broadly classified into two types: tumor suppressor genes and oncogenes. A tumor suppressor gene leads to cancer when there is a decrease of expression of the corresponding protein; an oncogene leads to cancer when there is an increase of expression of the corresponding protein. Our aim is to infer from the gains and losses of chromosomes and chromosome arms, which chromosomal regions are most likely to harbor important genes for tumor initiation, and which may be important for progression. We work with gain and loss information for regions of chromosomes, since this kind of a cytogenetic study provides a survey of the entire human genome at once. Furthermore, chromosomal alterations are known to

[1]Department of Mathematics, Rutgers University, Piscataway, New Jersey.
[2]Institute of Pathology, University of Basel, Basel, Switzerland.
[3]Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland.
[4]Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, California.
[5]Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda and Baltimore, Maryland.

lead to inactivation of tumor suppressor genes (by loss of chromosomal regions), or to oncogene activation (by gain of chromosomal regions).

Chromosomal abnormalities in cancer were first discovered in a form of leukemia (Nowell, 1976). During the past 20 years many forms of leukemia and lymphoma have been found to be associated with specific alterations. These studies have laid the foundations for contemporary cancer genetics. In solid forms of cancer, unlike leukemia and lymphoma, the study of chromosomal alterations has been difficult. An important problem in solid tumors is that once a set of critical genetic alterations develops, the cancer cell goes "out of control" and starts to accumulate seemingly random alterations. Since solid tumor samples often contain over a dozen chromosomal alterations it has proven difficult to identify the primary disease-causing events.

We focus on data collected by one important laboratory technique *comparative genome hybridization (CGH)*. CGH allows detection of all significant gains and losses in a single experiment per tumor (Kallioniemi *et al.*, 1992). CGH has opened interesting possibilities for understanding *oncogenesis*, the process whereby cancers form. CGH is based on the fact that each chromosomal region of a healthy cell has two copies of its DNA in a cell. Deviations from this normal level are called *copy number aberrations,* or *CNAs*. Significant deviations, up or down, in a cancerous cell may signal that the region contains genes that play a role in the development of cancer.

The copy number of chromosomal regions can be measured in the laboratory by measuring fluorescence on chromosomes where fluorescently labeled tumor DNA and normal DNA have been allowed to bond together. Comparing with the corresponding numbers in a healthy cell, one can obtain a list of CNAs. This is done by calculating for each region the ratio of copy numbers between healthy and cancerous cells (by comparing fluorescence of two colors), and selecting those regions for which this ratio is outside a range defined by a lower and an upper threshold, $r^- < 1 < r^+$, which take into account natural fluctuations and experimental error. An example of a successful application of CGH is the localization to chromosome 19 of a gene for Peutz-Jeghers syndrome (Hemminki *et al.*, 1997), a disease associated with precancerous intestinal polyps. For a survey of CGH and its applications, see Forozan *et al.* (1997).

Once we have the list of the copy number aberrations (CNAs) of a particular tumor, we can think of it as a set of genetic events that took place in some unknown order. It is believed that these events do not occur in a random fashion, but they are the result of some unknown flow of causality. That is, once an event occurs, it increases the probability of other events occurring, and so on. In some cases, the connection between one event and the next will be specific and directly causal, while in other cases the later event occurs seemingly at random because of the basic genetic instability in a tumor cell. A survey of many CGH studies shows some consistent patterns as to which CNAs occur most often in some types of tumors (Forozan *et al.*, 1997). By studying the sets of CNAs from many patients with the same kind of tumor, we may discover common patterns, ultimately we would like to infer the pathways whereby *oncogenesis* (the stochastic process in the genome which ultimately leads to cancer) proceeds. Models for tumor progression pathways would be of obvious value to the early diagnosis and treatment of cancer.

There has already been important work in this direction, notably the study of colorectal cancer by Vogelstein *et al.* (1988). They inferred from a variety of types of data that the progression of colorectal cancer can be described by a *chain* of four genetic events, three of which are CNAs. When the first of these events occurs, the chance of the second event occurring increases, and when the second event occurs the chances of the third increase, and so on. These events are irreversible, in that once an event occurs it is never undone in the future. The presence of all four events appears to be an indicator of colorectal cancer. While the path model suggests a most likely order of occurrence, colorectal cancer is really associated with the accumulation of the genetic changes on the path (Vogelstein *et al.*, 1988).

The aim of the present paper is to build on the above-mentioned work on inferring the oncogenesis process from CNA data, and to develop its mathematical and computational foundations. Previous analyses of CGH data (Kuukasjärvi *et al.*, 1997) suggest that *chain* or *path* models similar to that in Vogelstein *et al.* (1988) do not suffice to capture oncogenesis as suggested for colorectal cancer.

Therefore, we consider a more general *tree-like* model of oncogenesis. In other words, while Vogelstein *et al.* (1988) restrict the causal sequence of genetic events causing cancer to straight-line chains such as that shown in Figure 1A, we allow a much more general family of models that branch like *trees* (Figure 1B,C). By extending the set of models that we can handle, we increase our chances of correctly inferring how oncogenesis proceeds. Our models explicitly assume that the events may not be independent, in contrast to the work of Newton *et al.* (1994) who did a statistical analysis of a small set of chromosome loss data collected by cytogenetic methods. A correct model of oncogenesis would be valuable in suggesting narrow chromosome regions in which to focus subsequent experimental effort, as well as the relative timing of such efforts.

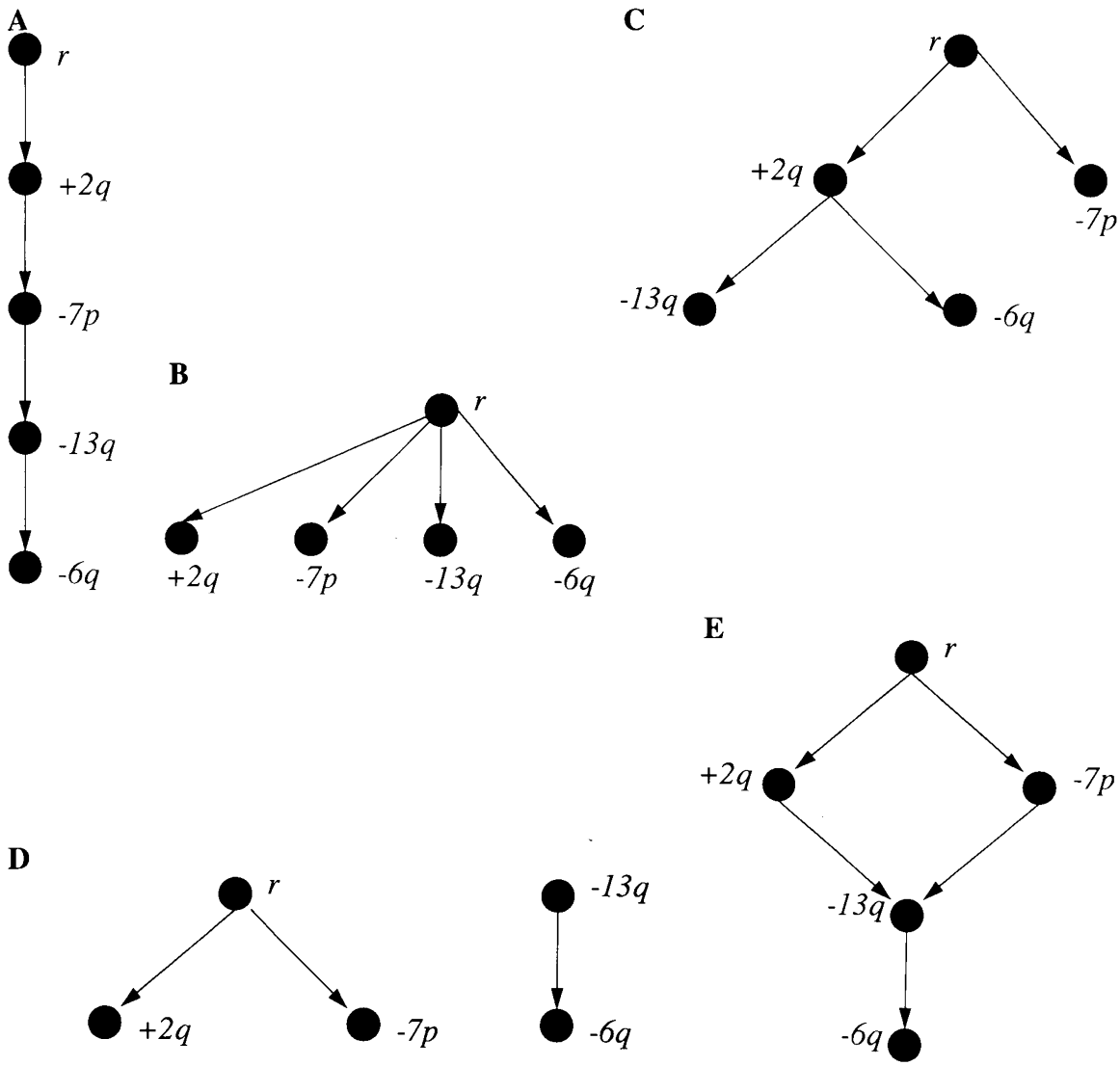

FIG. 1. Graph structures of some possible oncogenetic models. Part **A** is a path; part **B** is a star; part **C** is a tree that is neither a path nor a star; part **D** shows a two-component forest; and part **E** shows a directed acyclic graph that is not a tree. In this paper, we allow structures A–C, but we disallow E. In models such as D, we consider only the component with vertex r.

In Section 2 of this paper, we develop a mathematical theory of the oncogenesis process for tree-like models. The mathematical theory borrows from Markov processes and percolation theory. We distinguish between two styles of models, one akin to network reliability and the other involving time. We also define a statistic that may be valuable in distinguishing between events that occur early from those that occur late in oncogenesis.

In Section 3, we develop a methodology for inferring an oncogenesis tree from CNA data. Our methodology involves selecting a set of most relevant events, and then assigning to each pair of such events a weight related to the probabilities of joint or individual occurrence of these events. From these weights, we can recover the optimum oncogenesis tree as a *maximum-weight branching* corresponding to these weights. We prove that, under mild assumptions, our algorithm discovers the correct tree model of oncogenesis, if supplied with an appropriate amount of data. Our analysis can be made quantitative to yield an estimation of the amount of data that would suffice for correct inference with high probability. Sample size estimation is important because data collection in oncology is costly.

In Section 4, we use our methodology to analyze CNA data for renal cancer from the laboratory of H.M. We use the efficient algorithm suggested by our method to find a tree model that appears to capture the oncogenesis process for one category of renal cancer. We conclude in Section 5 with directions for future work.

2. ONCOGENETIC TREES

The result of a CGH test is a set of genetic events (in particular, a set of CNAs). A series of CGH tests results in a family of sets of CNAs, that is to say, a set of samples from a *probabilistic distribution* over all possible sets of genetic events. A model of oncogenesis should then be a random process which generates such sets of genetic events, and therefore defines a distribution over sets of genetic events. In this section we propose several such models, most notably *tree-based models.*

Let us fix a finite set V of possible genetic events. In practice, the genetic events are reported from the laboratory as a set of chromosome intervals gained or lost in each tumor. Human chromosomes are named/numbered 1 to 22 and X and Y (only in males). For our studies, we have ignored Y because it is small and some of the data sets are female breast cancer samples. All the chromosomes have a long arm, denoted q. All the chromosomes except 13, 14, 15, 21, and 22 have significant genetic material on the short arm denoted p. Studies of chromosomes have long relied on the fact that different intervals, called bands, along the chromosome, are colored differently after applying a stain. When two tumors have CNAs spanning approximately the same bands, we found it impossible to decide whether these reflect the same genetic change or not. Therefore, in our data analysis we used only the 41 chromosome arms and gain/loss to distinguish events. As a result, V contains 82 possible CNAs, plus a root note r. It is possible for the same chromosome arm to have both a gain and a loss (at different bands) in the same tumor. A *(probability) distribution on* 2^V is a function p assigning to each subset of V a nonnegative real number, and satisfying $\sum_{S \subseteq V} p[S] = 1$.

We shall consider graphs whose vertices are the genetic events in V, and which define distributions on 2^V. A *rooted tree*, or simply *tree*, on V is a triple $T = (V, E, r)$, where $r \in V$ is a special vertex called the *root*, and E is a set of pairs of vertices such that (1) for each vertex $v \in V$ there is at most one edge $(u, v) \in E$ with v as its second component, (2) there is no edge (u, r) entering r, and (3) there is no *cycle*, that is, no sequence of edges in E of the form $((v_0, v_1), (v_1, v_2), \ldots, (v_{k-1}, v_k), (v_k, v_0))$. For example, the graphs depicted in Figure 1A–D are trees, while the graph in 1E is not. Notice that we allow trees to have disconnected components; for our analysis the interesting part of a tree will be the one reachable from r. There are two special kinds of trees, in some sense opposite extreme cases, that are of special interest in our application: A *path* is a tree with at most one edge leaving each vertex (Figure 1A). A *star* is a tree in which all edges leave the root (Figure 1B). Note that what we call "tree" is often called a *branching* in the optimization literature (Papadimitriou and Steiglitz, 1982), where the term "tree" is reserved for the undirected version. Throughout this paper we consistently use the term "tree" mean a directed rooted tree, possibly disconnected; however, for consistency with the literature, we refer to the algorithm we employ as the "maximum branching algorithm."

A *labeled tree* $T = (V, E, r, \alpha)$ is a rooted tree with a positive real number $\alpha(e) > 0$ on each edge $e \in E$. Labeled trees are useful as generators of distributions on 2^V. Suppose that we are given a labeled rooted tree $T = (V, E, r, \alpha)$, where for all $e \in E\, 0 < \alpha(e) \leq 1$. We call such a tree an *oncogenetic tree;* intuitively, we can think of $\alpha(e)$ as the probability that edge e is present, with the events "edge e is present" being independent. We can then carry out the following experiment: Create a subtree of T by including each edge e with probability $\alpha(e)$, independently; then consider the set $S \subseteq V$ of all vertices reachable from r. S is the outcome of the experiment. Thus, such an oncogenetic tree T generates a distribution P_T on 2^V, where for each $S \subseteq V$ we have

- if $r \in S$ and there is a subset $E' \subseteq E$ such that S is the set of all vertices reachable from r in the tree (V, E', r), then

$$p[S] = \prod_{e \in E'} \alpha(e) \cdot \prod_{(u,v) \in E, u \in S, v \notin S} (1 - \alpha((u, v)));$$

- otherwise, $p[S] = 0$.

In significant parts of the paper, we use the logarithm of a probability, rather than the probability itself as an edge weight.

Oncogenetic trees are a simple but rigorous model of oncogenesis. It is assumed that the causality between genetic events is tree-like, and that a genetic event causation of a genetic event by another is independent of other such causations. Both assumptions are questionable, and are made for modeling simplicity and economy. A further objection, namely that such models do not capture the false positives and negatives due to experimental inaccuracies in CGH is discussed in the next section. Even if the true causality of oncogenesis

is a much richer non–tree-directed acyclic graph (that is, even if there is confluence of causality), it is hoped that most of the causality flows may be low enough in probability, and thus there is a tree-like model that captures accurately enough the dominant factors in oncogenesis. Similarly, statistical dependence of genetic events may be approximated adequately by independent edges. In any event, trees are a far richer model than the path models used so far in this arena (Vogelstein *et al.*, 1988).

The simple oncogenetic tree model we have introduced fails to take into account *time*, a factor of obvious importance to oncogenesis. We next introduce a more elaborate model, which we call *timed oncogenetic trees*. A timed oncogenetic tree is a labeled tree $T = (V, E, r, \lambda)$, together with a distribution ϕ on the positive reals. Timed oncogenetic trees represent the following sampling process. First, for each edge e we draw a random variable $t(e)$, exponentially distributed with mean $\frac{1}{\lambda(e)}$. Second, we draw a real number t_{tot} from distribution ϕ. We include a vertex v in the outcome of the sampling if and only if there is a path P from r to v in T, and the sum of all $t(e)$'s over all edges e on this path is at most t_{tot}.

Thus, in a timed oncogenetic tree we assume that event r happens at time 0 (in the application of this model, event r will be an extraneous event signaling the beginning of the process). Once event u has happened, for each edge $(u, v) \in E$ event v is a Poisson event with rate λ. We select those events that have happened by time t_{tot}, where t_{tot} is drawn from a distribution ϕ which ideally reflects the time, relative to the oncogenetic process, at which a tumor is sampled from a patient.

2.1. *Discussion of the models*

The timed oncogenetic tree model is a more realistic model of oncogenesis than the timeless oncogenetic tree, in which non-intersecting branches of the tree are independent. However, the timeless model also merits study because (a) it is more tractable mathematically, and (b) it is reasonable to hope that, once the best timeless oncogenetic tree has been identified, its structure will contain important clues about the best timed oncogenetic tree. In fact, we can show that in the case of *paths* the two models are equivalent:

Theorem 2.1. *Let $T = (V, E, r, \lambda)$ be a timed oncogenetic tree of path topology with distribution ϕ. Then there is a timeless oncogenetic tree T' generating the same distribution.*

Proof. Let $T = (V, E, r, \lambda)$ be a timed oncogenetic tree with edges $(r, v_1), (v_1, v_2), \ldots (v_{n-1}, v_n)$. We sample first by choosing t_{tot} according to ϕ. Let t_i be the random variable that represents the time at which v_i occurs. For $v_i \in V$, define ϕ_i to be the distribution of $t_{\text{tot}} - t_i$, conditioned upon $t_i < t_{\text{tot}}$. Let X be the set of events that occur before t_{tot}.

Suppose $e_1 = (r, v_1)$ has weight λ_1. Conditioned on $t_{\text{tot}} = s$, the probability of $v_1 \in X$ is $1 - e^{-\lambda_1 s}$. Integrating over s using the distribution ϕ leads to

$$p_1 = P[v_1 \in X] = \int_0^\infty (1 - e^{-\lambda_1 s})\phi(s)\, ds.$$

We thus have a closed form for the edge probability for an edge incident to the root, in terms of the time distribution ϕ.

Let $e_2 = (v_1, v_2)$. Define $p_2 = P[x_2 \in X \mid x_1 \in X]$. Since $\phi_1(t)$ is the distribution of $t_{\text{tot}} - t_1$ conditioned on $v_1 \in X$,

$$p_2 = \int_0^\infty (1 - e^{-\lambda_2 s})\phi_1(s)\, ds.$$

This reasoning can be extended for the entire path. With weight λ_i for each i, we define the distribution ϕ_i successively as the distribution of $t_{\text{tot}} - t_i$ conditioned on $v_i \in X$. We then define $p_i = p(e_i)$ for $i > 1$ by

$$p_i = \int_0^\infty (1 - e^{-\lambda_i s})\phi_{i-1}(s)\, ds.$$

■

This result implies that our methods for reconstructing timeless oncogenetic trees in the next section will yield the correct timed oncogenetic tree when the underlying timed oncogenetic tree is of a path topology.

Our oncogenetic tree models borrow from the theory of *ascending Markov chains* and *percolation processes* (for example, see Fill and Pemantle, 1993; Richardson, 1973), known to be valuable as models of physical and

biological systems. The timeless oncogenetic tree model is partly inspired by the closely related *Cavender-Farris model* for evolution (Cavender, 1978; Farris, 1973; Neyman, 1971). A *Cavender-Farris tree* is also a rooted tree $T = (V, E, r, p)$ which is also used to generate a distribution on 2^V. First, r is in S. Then, for each edge $e = (v_i, v_j)$, v_j differs from v_i with respect to inclusion in S probability p_e. If T is of star topology, then the oncogenetic tree model and the Cavender-Farris tree model are equivalent.

Finally, some notation. All our tree models define distributions on 2^V. We shall denote the distribution induced by oncogenetic tree T (timeless or timed, the context will be clear) as P_T. Let P be such a distribution. We shall need to define various associated probabilities. For $v_i, v_j \in V$, define

- $p_i = \sum_{v_i \in Y} P(Y)$
- $p_{ij} = \sum_{\{v_i, v_j\} \subseteq Y} P(Y)$
- $p_{i\neg j} = \sum_{Y | v_i \in Y, v_j \notin Y} P(Y)$
- $p_{i|j} = \frac{p_{ij}}{p_j}$
- $p_{i|\neg j} = \frac{p_{i\neg j}}{1-p_j}$

3. THE RECONSTRUCTION PROBLEM

In this section, we address the following important question: *Given a set of CGH data, how can we find the oncogenetic tree, in either model, that best fits the data?* In its purest form, the problem is one of characterizing the class of distributions P_T, for T an oncogenetic tree. This is not difficult in the case T is either a path (the support of P_T immediately suggests T) or a star (the events should be independent), but very challenging in its generality. In practice it is further complicated by the fact that CGH data are known to contain false positives and negatives, corresponding to either experimental or observational errors, or genetic events irrelevant to the cancer under study. False negatives can be incorporated in our oncogenetic tree model by introducing, for each vertex v, a new vertex v' with intuitive meaning "genetic event v was observed," and an edge (v, v') with probability one minus that of the false negative. We see no simple way to incorporate false positives in our models.

In this section we develop a methodology for reconstructing oncogenetic trees from CGH data. We formally show that, in the absence of experimental errors, and unless the underlying tree has a certain anomaly analogous to the *long branch attraction* of Cavender-Farris trees (Felsenstein, 1978), with enough data our method will correctly reconstruct the oncogenetic tree. Our result leads to explicit estimates on the size of data sets that are needed for the reconstruction method to be reliable. It is an open problem, discussed in the last section, how our method fares in theory in the face of experimental errors and false negatives and positives. However, in the next section we apply it to CGH data for renal cancer, with results that appear to be very satisfactory.

3.1. Reconstruction by maximum branchings

Our approach to the reconstruction problem is the following: Based on the CGH data, we construct a *weight* between the genetic events, that is, we assign a real w_{ij} for each pair (i, j) of genetic events; w_{ij} is in general asymmetric. We then find the rooted tree whose total weight (the sum of the weights of all edges in the tree) is maximized. The key algorithmic ingredient for this is a classical result due to Edmonds (Edmonds, 1967; Karp, 1971; Tarjan, 1977; Chu and Liu, 1965) stating that the best rooted, directed tree can be found efficiently—in fact, in $O(n^2)$ time, where n is the number of genetic events. Notice that this is *not* the better-known minimum spanning tree problem, which relates to undirected graphs (Papadimitriou and Steiglitz, 1982). In the application of this method to real data, there is a preprocessing step in which we use a maximum clique heuristic to omit from further consideration certain genetic events whose occurrence in the CGH data appears not to be important.

3.1.1. Skewed trees. We need to state and justify some assumptions about the probabilities of (combinations) of events to cope with real data and to prove that our reconstruction algorithm works. First, we need an assumption about edge probabilities to ensure that events are distinguishable and recognizable. Suppose $e = (v_i, v_j)$ is an edge in an oncogenetic tree. Were we to allow $\alpha(e)$ to approach 1, the two events i and j might be indistinguishable. If $\alpha(e) \to 0$, it may impossible to observe i and j together in any sample of reasonable size. Furthermore, if v_i is a vertex unlikely to be in a sample set, then its low probability must be allowed for. Also, if there is another edge (v_i, v_k), we must observe a reasonable number of sample sets where v_j was included while v_k was excluded, and vice versa.

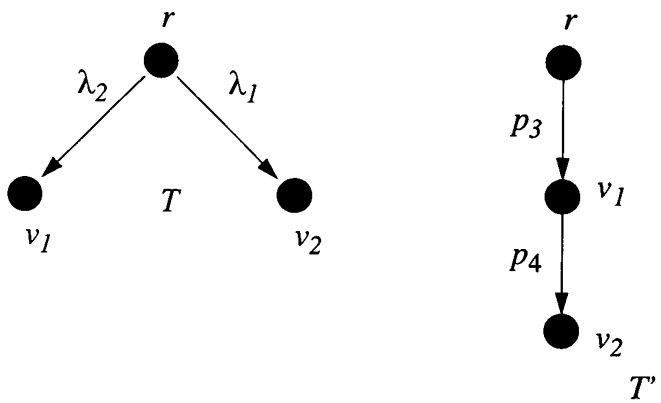

FIG. 2. Two timed oncogenetic trees that may be hard to distinguish.

Rigorously, we exclude all the above difficulties by choosing a constant $\epsilon > 0$ such that $p_i > \epsilon$ for each event i, and that for each pair of events i, j either $|p_i - p_j| > \epsilon$, or $p_i - p_{ij} > \epsilon$. The parameter ϵ is critical later when we estimate the sample size needed to infer the correct tree with high probability.

If we could sample an oncogenetic tree in a noise-free environment, then reconstruction could be done as follows. We could define the "ancestor" partial order $<_T$ by $v_i <_T v_j$ iff $p_{i|j} = 1$. We then construct a tree T which agrees with the partial order. With perfect sampling, this simple algorithm will return T when enough samples have been taken to distinguish each edge of T. In practice, however, this algorithm will almost certainly return a star because: (1) even when the model is correct, the CNAs will not always occur in the order of the model (Vogelstein *et al.*, 1988), (2) some of the CNAs that occur are truly random, and (3) the CGH experiment sometimes makes errors in reporting the CNAs. In the real data sets we have examined it is rare to find two events A and B such that both occur in many tumors and the occurrence of A strictly implies the occurrence of B.

When noise is allowed in the sampling process, edge probabilities must be large enough to distinguish the influence of edges from the influence of noise. Consider the two timed oncogenetic trees T and T' of Figure 2. We can choose the parameters, and the time distribution ϕ, so that the two distributions P_T and $P_{T'}$ are extremely close. For example, let $\delta > 0$ be chosen as a very small number. Suppose we use a two-state time distribution on the value of t, such that $t = t_1$ or t_2 each with probability $1/2$. We may choose edge weights for T, λ_1, λ_2 such that $P[v_1 \in X \mid t = t_1] = P[v_2 \in X \mid t = t_2] = 1/2$, while $P[v_1 \in X \mid t = t_2] \geq 1 - \delta$, while $P[v_2 \in X \mid t = t_1] \leq \delta$.

Compare this distribution P_T with the distribution $P_{T'}$ generated by taking $p_3 = 3/4$, $p_4 = 1/3$. (In defining $P_{T'}$, we take advantage of Theorem 2.1.) When noise is allowed, for small values of δ the two trees are virtually indistinguishable. Hence, no algorithm will successfully distinguish between the two trees with less than astronomically large data.

The difficulty with reconstructing T lies in the fact that the distribution does not adequately support the edge (r, v_2) instead of the edge (v_1, v_2). By allowing the fullest range of edge weights, we open the door to topological deception. This leads us to the following definition of extreme pathological cases of oncogenetic trees:

Definition: Let T be an oncogenetic tree, generating distribution P_T on V. We say T is *skewed* if there exists three distinct vertices $v_i, v_j, v_k \in V$, such that v_k is the least common ancestor of v_i, v_j, yet $p_{i|j} \geq p_{i \cup j|k}$.

Thus, we say T is skewed if the edge (v_i, v_j) has greater support than the two edges (v_k, v_i) and (v_k, v_j). In a timeless oncogenetic tree $p_{i|j} = p_{i|k} < p_{i \cup j|k}$, so *timeless oncogenetic trees are never skewed.*

3.1.2. The weight functional. Our next step is to define a *weight functional*, that is, a mapping from probability distributions over 2^V to real weights for the pairs in V^2. We shall use these weights to reconstruct the oncogenetic tree as the optimum branching (maximum-weight rooted tree) for these weights.

Intuitively, the weight w_{ij} should reflect the desirability of having j as a direct descendant of i in the tree. First, it should reflect the *likelihood ratio* for i and j occurring together, that is, $\frac{p_{ij}}{p_i p_j}$. There should also be an asymmetry in the weight as well to reflect which CNA is likely to occur first. If $p_i > p_j$, this means that event

i occurs more often that event j; hence, it is more advantageous to have an edge from i to j than vice-versa. This suggests the following functional:

$$w_{ij} = \frac{p_i}{p_i + p_j} \cdot \frac{p_{ij}}{p_i p_j}$$

To be able to prove that the reconstruction algorithm works, it turns out that the right choice is the *logarithm* of the above quantity:

$$w_{ij} = \log(p_{ij}) - \log(p_i + p_j) - \log(p_j) \tag{1}$$

If we were instead solving the minimum spanning tree problem, taking the logarithm or any other monotone increasing function, of the distances would not change the optimum tree. However, in the more elaborate case of branching, taking the logarithm may, in rare cases, change the optimum solution.

For weight functional (1), we can prove the following result:

Theorem 3.1. *Let T be a nonskewed oncogenetic tree (timeless or timed). Then the maximum branching over V with respect to the weights defined by equation (1) from the distribution P_T is precisely T.*

Since timeless oncogenetic trees are not skewed, and also since each timed oncogenetic path has an equivalent timeless oncogenetic path (by Theorem 2.1). Theorem 3.1 leads immediately to the following:

Corollary 3.2. *Let T be a timeless or path-like, timed oncogenetic tree. The maximum branching over V with respect to the weights defined by equation (1) from the distribution P_T correctly reconstructs T.*

To prove the theorem, let T be a nonskewed oncogenetic tree with root r, and let B be the maximum branching for the weights defined by equation (1) from the distribution P_T. We use three lemmas to show that $B = T$.

Lemma 3.3. *The root of B is r.*

For analyzing real data, r is an artificially added vertex, conceptually representing the cell at time 0 with no CNAs. As such, there is no reason to keep the branching algorithm from artificially making r to be the root. However, this complicates the algorithm, and since the simpler algorithm returns r as a root, this lemma is useful.

Proof. Suppose not. Let v_i be the root of B, and v_j be the parent of r in B (it may be the case that $i = j$). Consider the branching B' obtained by removing the arc (v_j, r) and adding the arc (r, v_i). Then

$$\begin{aligned} w(B') - w(B) &= w(r, v_i) - w(v_j, r) \\ &= -\log(1 + p_i) - \log(p_j) + \log(1 + p_j) > 0. \end{aligned} \tag{1}$$

This contradicts the maximality of B. ■

Let $<_T$ represent the ancestor relation in T.

Lemma 3.4. *Let $v_j \in V$, $v_j \neq r$. Let v_i be the parent of v_j in B. Then $v_i <_T v_j$.*

Proof. Suppose not. Choose v_j closest to r in T with parent in B not an ancestor in T. Let v_i be its parent in B. Let v_k be the least common ancestor of v_i, v_j in T. Note that by choice of v_j, in B, v_k and all of its ancestors have parents in B which are ancestors in T.

Consider the branching B' obtained by deleting the edge (v_i, v_j) and adding the edge (v_k, v_j). By choice of v_j, adding (v_k, v_j) will create no cycles, and thus B' is a valid branching. Observe that

$$w(B') - w(B) = w(v_k, v_j) - w(v_i, v_j).$$

Now $w(v_k, v_j) = -\log(p_k + p_j)$ because the other two terms cancel since the occurrence of event j implies the occurrence of event k. Also, $w(v_i, v_j) = \log(\frac{p_{ij}}{p_j(p_i + p_j)})$. Thus

$$w(v_k, v_j) - w(v_i, v_j) = \log\left(\frac{p_j(p_i + p_j)}{p_{ij}(p_j + p_k)}\right). \tag{2}$$

Thus, to show $w(B)$ is maximal, it suffices to show $\frac{p_i + p_j}{p_j + p_k} > \frac{p_{ij}}{p_j}$. Now $\frac{p_{ij}}{p_j} = p_{i|j}$, and since T does not exhibit long branch attraction,

$$p_{i|j} < p_{i \vee j|k} = p_{i|k} + p_{j|k} - p_{ij|k}.$$

Thus

$$p_{i|j} + p_{ij|k} < p_{i|k} + p_{j|k},$$

which leads to

$$\frac{p_{ij}}{p_j}(p_j + p_k) < p_i + p_j,$$

which is equivalent to the desired inequality $p_{i|j} < \frac{p_i + p_j}{p_j + p_k}$. Thus $w(B') > w(B)$, a contradiction. ■

Lemma 3.5. *For every $v \in V$, $v \neq r$, the parent of v in B is the parent of v in T.*

Proof. Suppose not. Let $v_j \in V$, with parent v_i. Suppose $v_k \neq v_i$ is the parent of v_j in B. Then by Lemma 3.4, $v_k <_T v_i <_T v_j$. Consider the branching B' obtained by deleting edge (v_k, v_j) and adding (v_i, v_j). (Note that Lemma 3.4 ensures that B' is a branching.) Since $w(v_k, v_j) = -\log(p_k + p_j)$ and $w(v_i, v_j) = -\log(p_i + p_j)$, and $p_i < p_k$, then

$$w(B') - w(B) = \log(p_k + p_j) - \log(p_i + p_j) > 0. \tag{3}$$

Thus $w(B') > w(B)$, a contradiction. ■

To finish the proof of Theorem 3.1, by Lemma 3.3 and Lemma 3.5, B is a spanning tree of V with root r corresponding to the universal event, such that for each $v \in V$, the parent of v in T is also its parent in B. Thus, $B = T$. ■

Theorem 3.1 applies when we know the precise probability distribution P_T. In practice, however, we know P_T through *samples*. It turns out that a slight modification of the proof of the theorem covers this case as well: Given a set of samples from P_T, calculate $\hat{p}_i$, and $\hat{p}_{ij}$, the observed values of p_i and p_{ij} for $v_i, v_j \in V$. and from these $\hat{w}(v_i, v_j) = \log(\hat{p}_{ij}) - \log(\hat{p}_j) - \log(\hat{p}_i + \hat{p}_j)$, the observed value for $w(v_i, v_j)$. Find the maximum branching B with respect to $\hat{w}$.

Theorem 3.6. *If T is a nonskewed oncogenetic tree, then with sufficiently many samples, the above algorithm will correctly construct the tree T with high probability.*

To verify $B = T$, we use the three lemmas above. We first analyze Lemmas 3.3 and 3.5 because they carry over to the probabilistic setting easily. Lemma 3.4 is where the nonskewness assumption is used in the proof. Therefore a more complicated estimation is needed to relate the number of samples needed for the Lemma to apply to a measure of how strongly the nonskewness assumption holds.

We first test that r is the root of B. By equation (1), r will be the root of B if there are no pairs of vertices v_i, v_j such that

$$\log \frac{(1 + \hat{p}_j)}{\hat{p}_j(1 + \hat{p}_i)} \leq 0.$$

This is immediate if $\hat{p}_i < 1$ for each event v_i. (If $p_i = 1$, we may group v_i with r at the root of the tree.)

We next show that the proof of Lemma 3.5 also carries through for observed probabilities. Let v_k be the parent of v_j in B. Suppose $v_i \neq v_k$ is the parent of v_j in T. Consider the branching B' obtained by deleting edge (v_i, v_j) and adding edge (v_k, v_j). Then $w(B') - w(B) = \log(\hat{p}_k + \hat{p}_j) - \log(\hat{p}_i + \hat{p}_j)$. Presuming that at least one sample is taken with v_k but not $v_i, \hat{p}_k > \hat{p}_i$, so $w(B') > w(B)$, which contradicts the maximality of the weight of B.

We now show that we may extend a version of Lemma 3.4 to the algorithm analysis. For $v_j \neq r$, we verify that if $(v_i, v_j) \in B$, then v_i is an ancestor of v_j. Suppose not.

Let v_j be the vertex closest to r which has been assigned a nonancestor in T to be its parent in $\hat{B}$. Let v_i be its parent in $\hat{B}$ and let v_k be the least common ancestor of v_i, v_j in T. Define $\epsilon_{i|j} = p_{i\vee j|k} - p_{i|j}$. Since T is not skew, $\epsilon_{i|j} > 0$. Let $\hat{B}'$ be the branching obtained by deleting edge (v_i, v_j) and adding edge (v_k, v_j).

Suppose the branching contains edges (v_i, v_j) where v_i is not an ancestor of v_j. Let v_j be the vertex closest to r which has been assigned a nonancestor in T to be its parent in $\hat{B}$. Let v_i be its parent in $\hat{B}$ and let v_k be the least common ancestor of v_i, v_j in T. Let $\hat{B}'$ be the branching obtained by deleting edge (v_i, v_j) and adding edge (v_k, v_j). Then, referring back to Equation (2),

$$\begin{aligned} w(\hat{B}') - w(\hat{B}) &= w(v_k, v_j) - w(v_i, v_j) \\ &= \log(\hat{p}_{jk}) - \log(\hat{p}_k + \hat{p}_j) - \log(\hat{p}_{ij}) + \log(\hat{p}_i + \hat{p}_j) \\ &= \log\left(\frac{\hat{p}_{i|k} + \hat{p}_{j|k}}{\hat{p}_{i|j}(1 + \hat{p}_{j|k})}\right). \end{aligned}$$

Let $\hat{\epsilon}_{i|j} = \hat{p}_{i\vee j|k} - \hat{p}_{i|j}$. This is the observed value of $\epsilon_{i|j}$. From the proof of Lemma 3.4, we know that if $\hat{\epsilon}_{i|j} > 0$, then $w(B') > w(B)$, a contradiction. Define $\delta_{i|j} = \hat{p}_{i|j} - p_{i|j}$, and $\delta_{i\vee j|k} = \hat{p}_{i\vee j|k} - p_{i\vee j|k}$. We will observe $\hat{\epsilon}_{i|j} > 0$ if $\delta_{i|j} + \delta_{i\vee j|k} < \epsilon_{i|j}$. Let $\epsilon = \min_{i,j} \epsilon_{i|j}$.

We use the Chernoff bound on $\delta_{i|j}$ and $\delta_{i\cup j|k}$ (Alon *et al.*, 1992). We express the Chernoff bound in terms of a parameter u to be chosen shortly, the number of samples N, and the smallest probability among the events we are considering $p_{\min}$. The Chernoff bound states that

$$P\left[\delta_{i|j} > \frac{u}{\sqrt{N p_{\min}}}\right] < e^{-u^2/2}.$$

If we let $u^2 = 8 \ln n$, the right-hand side of the above equals n^{-4}. Thus, for this value of u, we observe that

$$P\left[\max_{i,j} \delta_{i|j} > \frac{u}{\sqrt{N p_{\min}}}\right] < \frac{1}{2n^2},$$

where the maximum is taken over the $\binom{n}{2}$ pairs $\{v_i, v_j\}$. A similar argument shows that

$$P\left[\max_{i,j,k} \delta_{i\cup j|k} > \frac{u}{\sqrt{N p_{\min}}}\right] < \frac{1}{2n^2},$$

where the maximum is taken over the set of triples (i, j, k) such that $v_k = lca(v_i, v_j)$. Setting $\epsilon = \frac{u}{N p_{\min}}$, we solve for

$$N = \frac{u^2}{\epsilon^2 p_{\min}} = \frac{8 \ln n}{\epsilon^2 p_{\min}}.$$

This proves a quantitative version of Theorem 3.6

Theorem 3.7. *If T is a tree with n vertices (not including the root r), and $p_{\min}$ is the minimum probability of observing any event, and ϵ is defined as above, then with $N = \frac{8 \ln n}{\epsilon^2 p_{\min}}$ samples of P_T, the probability that the algorithm returns a false edge is less than $1/n^2$.*

A realistic size for a tree might be five vertices, not including the artificial root r; this is one more vertex than in the colorectal cancer path model. Based on the data sets we have seen $p_{\min} = 0.2$ is plausible if one uses only the most common events to build the tree. It is not possible to directly estimate the most important parameter ϵ. If $\epsilon = 0.1$, it would take a number of samples near 6,400 to get every tree edge correct with probability of $>24/25$. The error rate of $1/25$ is appropriate as it loosely corresponds to a p value of 0.04 in a field where 0.05 is often used as a cutoff. However, the sample size 6,400 is nearly two orders of magnitude

too large to reasonably expect from CGH studies using current technology. Nevertheless, one would hope that the tree inferred from a smaller sample still has a high probability having most edges correct.

3.2. *The size statistic*

In this subsection, we define a statistic that is quite useful for establishing the relative temporal order of genetic events in oncogenesis. As we shall see, in the case of path trees, this simple statistic is sufficient for reconstruction.

Let $T = (V, E, r, \lambda)$ be an oncogenetic tree (timeless or timed) with root r and vertex set V. Let s be a positive, real-valued function on $V(T)$ for a rooted tree T. We say s *preserves the order of* T if $v_i <_T v_j$ implies that $s(v_i) \leq s(v_j)$.

For $v \in V$, define the *size* of v to be $s(x) = E_{P_T}[|X| \mid v \in X]$, i.e., the expected number of CNAs in tumors that contain v. Intuitively, the later v occurs the more likely that it will occur together with more CNAs.

Theorem 3.8. *s as defined above preserves tree order.*

Proof. Since $v_i <_T v_j$, if $v_j \in X$, then $v_i \in X$. Let $p = P[v_j \in X \mid v_i \in X]$. We expand

$$s(v_i) = pE[|X| : v_i \in X \wedge v_j \in X] + (1 - p)E[|X| : v_i \in X \wedge v_j \notin X].$$

To show $s(v_i) < s(v_j)$, it suffices to show

$$E[|X| : v_i \in X \wedge v_j \notin X] < E[|X| : v_j \in X],$$

which is true because v_i precedes v_j on the path.

The statistic s is simple, and correctly reconstructs the oncogenetic tree in the special case in which the tree has a path topology. In general, it may give valuable information on the stage of cancer at which the given genetic event occurs. The survey article (Forozan *et al.*, 1997) highlights the importance of using CGH data to identify which CNAs occur early and late in tumor progression; the size statistic gives a simple rigorous way to make hypotheses, to be tested in the laboratory, about early/late CNAs.

4. AN ONCOGENETIC TREE FOR RENAL CANCER

In this section we describe how to use our methods on a set of 117 cases of clear cell renal cell carcinoma from the laboratory of H.M. that was collected using CGH as described in (Jiang *et al.*, 1998; Moch *et al.*, 1996). Kidney cancer is known to be quite heterogeneous in its histology and its genetic origin (for a recent review, see Erlandsson, 1998). "Clear cell" describes one histological category of nonpapillary renal carcinoma. Kidney cancer has both familial and sporadic forms. The familial forms can, for example, be caused by a germ-line defect (i.e., inherited at birth and present in all cells) in a tumor-suppressor gene. This means typically that the copy of the gene on one of two homologous pairs of chromosomes is defective at conception, and the cancer occurs if the other copy gets lost in a renal cell. Most renal cancers (>90%; Motzer *et al.*, 1996) are sporadic cases where two mutations occur only after birth and only in some cells. The gene, which is responsible for the development of clear cell renal cell carcinoma associated with the rare von Hippel–Lindau syndrome has been identified on chromosome arm 3p (Latif *et al.*, 1993). About 70–80% of sporadic clear cell renal cell carcinomas have a loss of the von Hippel–Lindau gene on chromosome arm 3p (Gnarra *et al.*, 1994; Moch *et al.*, 1998). Another important example of a type of kidney cancer for which causative genes are known is papillary renal cell carcinoma for which one specific gene has been identified on chromosome arm 1q (Sidhar *et al.*, 1996; Westerman *et al.*, 1996) and for which the oncogene MET (involved in many types of cancers) on arm 7q has also been implicated (Schmidt *et al.*, 1997).

Studies of the role of the von Hippel–Lindau gene in renal cancer suggest that it should be affected early in the oncogenesis process and that a loss of this gene alone may not be sufficient to cause renal cancer. One example of such a study is that of Thrash-Bingham *et al.* (1995) who used a different laboratory technique to look for only losses in 33 renal cell carcinomas, of which 13 cases are clear cell renal carcinomas like the 117 we considered. Among those 13 cases Thrash-Bingham *et al.* (1995) observed five losses on 3p, four

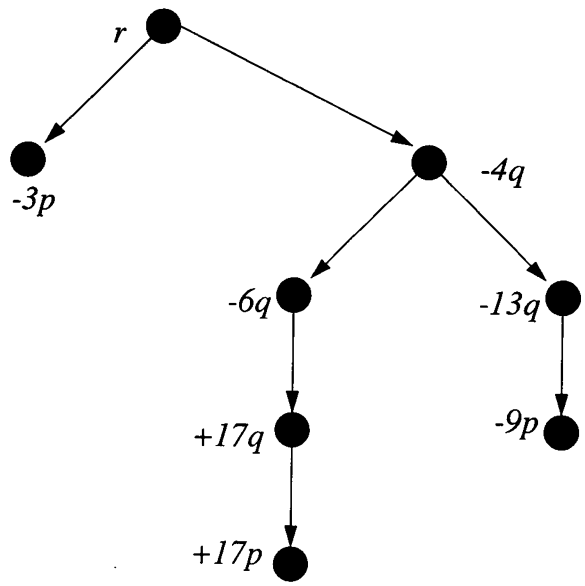

FIG. 3. Graph structure of our oncogenetic model for clear cell renal cell carcinoma.

losses on 14q, three losses on 8p, two losses on 6q, and no other repeated losses. In their overall study, they concluded that loss of 3p was associated with certain specific other losses, but this does not follow for the 13 clear cell cases alone.

Before applying the branching algorithm it is necessary to select a small set of events to work with that appear to be most relevant. A model involving all 82 CNA events is clearly not appropriate. Because cancer inherently involves genetic instability and because CGH does have some false positives, many of the possible 82 events will show up in a small percentage of the tumors. To select events that seem most relevant we used a *clique* heuristic on the weighted graph of all 82 events. An alternative method is suggested by Brodeur *et al.* (1982), but requires extensive simulations and good prior knowledge of the probability of each event. A clique is a subgraph in which all pairs of vertices are connected. For the clique computation the weight of an undirected edge is the sum of the two directed-edge weights defined for the branching computation. We restricted attention to pairs of events that occurred at least 5 times together in the same tumor. We then found a maximum-weight clique of size 7. The choice of size 7 was due to time concerns with the clique selection routine. It is important to select enough events to get an interesting model, while small enough to include events that occur often enough to be very likely to be relevant to renal cancer. Our clique of size 7 includes: -4q, -13q, -3p, $+17$p, $+17$q, -6q, -9p. We then applied the maximum-weight branching algorithm to the directed graph of these 7 vertices and the special vertex r. The edge structure of the best tree is shown in Figure 3.

Our model is consistent with the established theory that a loss on 3p is an early important event for clear cell renal carcinoma, and suggests that it is not causatively associated with specific other gains or losses. It is encouraging to see that changes in on 1q and 7q associated with papillary renal carcinomas do not show up in our model. It is also encouraging to see that a loss on 6q, which was one of the three other multiply-occurring events in (Thrash-Bingham *et al.*, 1995) does get selected in our much larger data set. Our model suggests that a loss on 4q is an important early event for clear cell renal carcinomas; a loss on 4q was observed in one case in (Thrash-Bingham *et al.*, 1995). This is also weakly supported by a summary of 3 CGH studies on renal cancer in (Forozan *et al.*, 1997) that showed a disproportionate number of cases with a loss on 4q, but this summary mixes together different types of renal cancer.

5. DISCUSSION

CGH is a powerful tool to explore the genetic defects underlying cancer. Since many cancers have heterogeneous genetic causes, mathematical modeling should help elucidate which CGH-detected aberrations may be causative and/or early prognosticators of tumor progression. The cancer genetics community has been quite excited by the path model for colorectal cancer (Fearon and Vogelstein, 1990), although analysis of CGH

data for colorectal cancer suggest that it may be oversimplified (Ried *et al.*, 1996). Moreover, other types of cancer do not seem to have useful path models.

We proposed two classes of tree models to relate CGH data to tumor progression. We described an algorithm to infer trees from CGH data and we proved that under one assumption about the tree structure, the algorithm infers the correct tree. We also estimated the sample size needed to get every edge in the tree correct with high probability. We partially validated our approach on a renal cancer data set. Not enough is known experimentally about tumor progression and CGH-detectable aberrations to validate our models completely. Along these lines it should reiterated that the usage of time in our models reflects the inherent "ascertainment bias" that only tumors, not healthy cells, are sampled, and we know little about how long the tumor has existed when it is sampled. One could instead imagine that the distribution of the sampling times may depend on the state of the cell, but such a model seems more appropriate for cell lines grown in a laboratory where repeated sampling at designated intervals is feasible. The software used in Section 4 is available by sending email to either **r.desper@dfkz-heidelberg.de** or **schaffer@helix.nih.gov**.

Our models imply that tumors of the type in the sample progress by starting at the root (no events) and then possibly adding any event whose parent in the tree has already been reached. They also imply that when certain combinations of vertices are reached, the abnormal cell would be sufficiently abnormal to be considered cancerous. The information we are most interested in, however, is which events are children of the root (indicative of early events) and which events are parent-child (possible causal relationship). It is extremely hard to figure out the cause and effect relationships in the laboratory, so a plausible model should be quite helpful to cancer geneticists.

It is not the case that reaching a single leaf implies cancer, as illustrated by our renal cancer model where the important event 3p- is a leaf next to the root. Current studies suggest that 3p- alone is insufficient for a renal cell to be cancerous. Our model suggests that for the type of renal cancer sampled, none of the 6 other events in the model occurs consistently after 3p-, although some of the tumors with 3p- do have many other gains or losses. We cannot easily infer which reached vertex sets are (not) cancerous because CGH data sets typically do not contain samples of cells that are abnormal, but not yet cancerous. The path model for colorectal cancer was in part validated because the genetic events along the path were correlated with clearly visible, precancerous abnormalities on the colon. An oncologist would have no easy way to sample precancerous, abnormal cells, except when the abnormality is visible. There are some other cancers, such as melanoma, that have this helpful feature, but many common cancers such as breast and prostate cancer do not.

Our method of modeling was partly inspired by Cavender-Farris evolutionary trees, especially some understanding of what conditions make those trees hard to infer correctly (Felsenstein, 1978). A recent poster abstract by Buetow *et al.* (1998) makes a direct connection between phylogenetic inference and tumor modeling. Buetow *et al.* (1998) used a standard phylogeny software package to build a tree of liver cancer samples, in which each leaf represents a distinct tumor and each internal vertex represents events whose presence/absence distinguishes the subtrees below. The purpose was to classify the tumors by their genetic characteristics. They used loss of heterozygosity of tandem-repeat DNA marker loci as the events. Are there other ways to mathematically exploit the analogy between evolving species and evolving tumors?

Among the other problems we left open are:

- How can one incorporate false positives into a tree model?
- Is it possible to make the clique heuristic rigorous or to find some other rigorous method for finding the most important CNA events other than that of (Brodeur *et al.*, 1982)?
- For what other plausible weight functions does the branching construction infer the correct tree with high probability?
- Suppose the best branching of any topology has weight W, while the star has weight $S < W$ and the best path has weight $P < W$. Can these differences in weight be converted into a *p value* for the hypothesis that the best model is not a star/path?
- What is the computational complexity of the tree reconstruction problem: Given a probability distribution P on 2^V, find a tree T such that some distance measure between P and $P(T)$ is minimized?

ACKNOWLEDGMENTS

Special thanks go to Martin Farach-Colton for his assistance as R.D.'s Ph.D. advisor at Rutgers University. R.D.'s work at Rutgers University was supported in part by NSF grant 94-12594. R.D.'s work was done partly

while he was a summer student intern at N. I. H. Thanks to Richard Simon for pointing out the relevance of Brodeur *et al.* (1982). The work of F.J. and H.M. was supported in part by grant number 31-50752.97 from the Swiss National Science Foundation. The work of C.H.P was supported in part by NSF grant CCR-9626361.

REFERENCES

Alon, N., Spencer, J.H., and Erdös, P. 1992. *The Probabilistic Method.* John Wiley & Sons, New York.

Brodeur, G.M., Tsiatis, A.A., Williams, D.L., Luthardt, F.W., and Green, A.A. 1982. Statistical analysis of cytogenetic abnormalities in human cancer cells. *Cancer Genet. Cytogenet.* 7, 137–152.

Buetow, K.H., Edmonson, M.N., Shen, F.M., Chen, G.C., London, W.T., and McGlynn, K.A. 1998. Identification of molecular heterogeneity in HCC using STRPs and tree-building algorithms. *Am. J. Hum. Genet.* 63, A336 (abst).

Cavender, J.A. 1978. Taxonomy with confidence. *Math. Biosci.* 40, 271–280.

Chu, Y.J., and Liu, T.H. 1965. On the shortest arborescence of a directed graph. *Sci. Sinica* 14, 1396–1400.

Edmonds, J. 1967. Optimum branchings. *J. Res. Nat. Bur. Stand.* 71B, 233–240.

Erlandsson, R. 1998. Molecular genetics of renal cell carcinoma. *Cancer Genet. Cytogenet.* 104, 1–18.

Farris, J.S. 1973. A probability model for inferring evolutionary trees. *Syst. Zool.* 22, 250–256.

Fearon, E., and Vogelstein, B. 1990. A genetic model for colorectal tumorigenesis. *Cell* 61, 759–767.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410.

Fill, J.A., and Pemantle, R. 1993. Percolation, first-passage percolation, and covering times for Richardson's model on the n-cube. *Ann. Appl. Prob.* 3, 593–629.

Forozan, F., Karhu, R., Kononen, J., Kallioniemi, A., and Kallioniemi, O.-P. 1997. Genome screening by comparative genome hybridization. *Trends Genet.* 13, 405–409.

Gnarra, J.R., Tory, K., Weng, Y., *et al.* 1994. Mutations of the VHL tumour suppressor gene in renal carcinoma. *Nat. Genet.* 7, 85–90.

Hemminki, A., Tomlinson, I., Markie, D., *et al.* 1997. Localization of a susceptibility locus for Peutz-Jeghers syndrome to 19p using comparative genomic hybridization and targeted linkage analysis. *Nat. Genet.* 15, 87–90.

Jiang, F., Richter, J., Schraml, P., *et al.* 1998. Chromosomal imbalances in papillary renal cell carcinoma: genetic differences between histological subtypes. *Am. J. Pathol.* 153, 1467–1473.

Kallioniemi, A., Kallioniemi, O.P., Sudar, D., *et al.* 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* 258, 818–821.

Karp, R.M. 1971. A simple derivation of Edmonds' algorithm for optimum branching. *Networks* 1, 265–272.

Kuukasjärvi, T., Karhu, R., Tanner, M., *et al.* 1997. Genetic heterogeneity and clonal evolution underlying development of asynchronous metastasis in human breast cancer. *Cancer Res.* 57, 1597–1604.

Latif, F., Tory, K., Gnarra, J., *et al.* 1993. Identification of the von Hippel–Lindau disease tumor suppressor gene. *Science* 260, 1317–1320.

Moch, H., Presti, J.C., Sauter, G., *et al.* 1996. Genetic aberrations detected by comparative genomic hybridization are associated with clinical outcome in renal cell carcinoma. *Cancer Res.* 56, 27–30.

Moch, H., Schraml, P., Bubendorf, L., *et al.* 1998. Intratumoral heterogeneity of von Hippel–Lindau gene deletions in renal cell carcinoma detected by fluorescence *in situ* hybridization. *Cancer Res.* 58, 2304–2309.

Motzer, R.J., Bander, N.H., and Nanus, D.M. 1996. Renal cell carcinoma. *N. Engl. J. Med.* 335, 865–875.

Newton, M.A., Wu, S.-Q., and Reznikoff, C.A. 1994. Assessing the significance of chromosome-loss data: where are suppressor genes for bladder cancer? *Stat. Med.* 13, 839–858.

Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems, 1–27. *In* Gupta, S.S., and Yackel, J., eds., *Statistical Decision Theory and Related Topics*, Academic Press, New York.

Nowell, P.C. 1976. The clonal evolution of tumor cell populations. *Science* 194, 23–28.

Papadimitriou, C.H., and Steiglitz, K. 1982. *Combinatorial Optimization.* Prentice Hall, Englewood Cliffs, NJ.

Richardson, D. 1973. Random growth in a tesselation. *Proc. Camb. Phil. Soc.* 74, 515–528.

Ried, T., Knutzen, R., Steinbeck, R., *et al.* 1996. Comparative genomic hybridization reveals a specific pattern of chromosomal gains and losses during the genesis of colorectal tumors. *Genes Chromosomes Cancer* 15, 234–245.

Schmidt, L., Duh, F.-M., Chen, F., *et al.* 1997. Germline and somatic mutations in the tyrosine kinase domain of the MET proto-oncogene in papillary renal carcinomas. *Nat. Genet.* 16, 68–73.

Sidhar, S.K., Clark, J., Gill, S., *et al.* 1996. The t(X;1)(p11.2;q21.2) translocation in papillary renal cell carcinoma fuses a novel gene PRCC to the TFE3 transcription factor gene. *Hum. Mol. Genet.* 5, 1333–1338.

Tarjan, R.E. 1977. Finding optimum branchings. *Networks* 7, 25–35.

Thrash-Bingham, C.A., Salazar, H., Freed, J.J., Greenberg, R.E., and Tartof, K.D. 1995. Genomic alterations and instabilities in renal cell carcinomas and their relationship to tumor pathology. *Cancer Res.* 55, 6189–6195.

Tirkkonen, M., Johannson, O., Agnarsson, B., *et al.* 1997. Distinct somatic genetic changes associated with tumor progression in carriers of *BRCA1* and *BRCA2* germ-line mutations. *Cancer Res.* 57, 1222–1227.

Vogelstein, B., Fearon, E., Hamilton, S., *et al.* 1988. Genetic alterations during colorectal-tumor development. *N. Engl. J. Med.* 319, 525–532.

Westerman, M.A.J., Wilbrink, M., and Geurts van Kessel, A. 1996. Fusion of the transcription factor TFE3 gene to a novel gene, PRCC, in t(X;1)(p11;q21)-positive papillary renal cell carcinomas. 1996. *Proc. Natl. Acad. Sci.* U.S.A. 93, 15294–15298.

Address reprint requests to:
Alejandro A. Schäffer
NCBI/NIH
Building 38A, Room 8N805
8600 Rockville Pike
Bethesda, MD 20894

Schaffer@helix.nih.gov

Received August 16, 1998; accepted December 13, 1998.